RESEARCH ARTICLE

Method of Genetic Disease Gene Locus Analysis Based on Apriori Algorithm

Periyasamy Mondal[1], Neeraj Acharya[2]

[1] Research Institute for Medical Genetics and Caring

[2] Moscow Centre of Chemical and Technology

Abstract: Many phenotypic traits of the human body and their susceptibility to drugs and diseases may be associated with some genetic loci or associated with genes that contain multiple loci. Whole genome association analysis (GWAS) is a hot spot in the analysis of genetic disease loci. In this paper, a Apriori algorithm based on Apriori algorithm is proposed for the analysis of genetic disease loci in genetic diseases. The experimental results show that the algorithm can effectively find the genetic disease gene loci.

Keywords: Apriori algorithm; genetic disease; gene locus; R language

1. Introduction

The human genetic code is contained in the DNA double helix long chain molecule by the human body, and the gene is the DNA long chain has the genetic effect some fragments. In the vast number of base pairs (or corresponding deoxynucleotides) that make up DNA, there are a number of single nucleotides at specific locations that frequently mutate to cause DNA polymorphism, which we call loci. A large number of studies have shown that many phenotypic differences in the human body, as well as susceptibility to drugs and diseases, may be associated with certain loci, or with genes containing multiple loci. Therefore, locating loci associated with traits or diseases in chromosomes or genes can help researchers understand the genetic mechanism of traits and some diseases, and can also enable people to intervene in the pathogenic loci to prevent the occurrence of some genetic diseases. In recent years, with the implementation of the Human Genome Project and the Genome Haploid Mapping Project, a large number of genetic variations associated with human traits or complex diseases have been identified and identified by genome-wide association analysis (GWAS), which provides important information for further understanding the genetic characteristics of human complex diseases. Clues.[1] In this paper, aiming at the problem of genetic disease gene pathogenic locus detection, a method of genetic disease gene pathogenic locus analysis based on Apriori algorithm is proposed.

2. Problem description

Table 1. 6 samples of genome wide data information (3 patients, 3 healthy persons)

A genome-wide approach to identifying a disease site or gene is employed by recruiting a large number of volunteers (samples), including people with a genetic disorder and healthy people, usually one for a patient and zero for a healthy person. For each sample, the base (A, T, C, G) encoding method is used to obtain the information of each site (be-	Sample health status	Chromosome fragment locus name and locus allele information			
		rs100015	rs56341	...	rs21132
1	1	TT	CA	...	GT
2	0	TT	CC	...	GG
3	1	TC	CC	...	GG
4	1	TC	CA	...	GG
5	0	CC	CC	...	GG
6	0	TT	CC	...	GG

3. Method

doi: 10.18063/gds.v2i1.

# of genetic disease loci analysis based on Apriori algorithm

3.1 Single locus pathogenic analysis

Because we need to analyze the association between a single site and whether it is pathogenic or not, the data is relatively small, so we only analyze the results according to the most basic Apriori algorithm. And the association between the coding information of a sample's locus on a potentially pathogenic chromosome fragment and the information of the sample's genetic disease. To find the most probable one or several pathogenic sites. Based on the base pair coding, we compared the undiagnosed samples with the diseased samples to find the base pair distribution of the site, so as to lock all the relevant pathogenic sites.

3.2 Pathogenic analysis of single gene (multiple loci)

Through the improvement of the algorithm, the influence of all the loci contained in each gene on genetic disease A is analyzed, and the correlation between each gene and disease A is analyzed. Because the frequent itemsets in Apriori algorithm contain one or more data, the association between genetic diseases and genes can be expressed by a complete set or a subset of the loci contained in the genes, so in order to show more clearly the true impact of each gene on genetic disease, we choose to get the frequency. Complex item sets are added from one to more lift values. To analyze the effect of all the loci contained in each gene on genetic diseases, and to analyze the association between each gene and disease.

3.3 Method of genetic disease loci analysis based on Apriori algorithm

3.3.1 The main idea of Apriori algorithm is to find out all the frequency sets, which appear at least as frequently as the predefined minimum support

Then the strong association rules are generated from the frequency sets. These rules must satisfy the minimum support and minimum credibility. Then, the desired rule is generated using the frequency set found in step 1, and all the rules that contain only the items of the set are generated, with only one item on the right of each rule. The definition of the intermediate rule is used here. Once these rules are generated, only rules larger than the minimum confidence given by the user are left. In order to generate all frequency sets, recursive methods are used.[2]

3.3.2 Some concepts and definitions

(1) Transaction Database: D, which stores two dimensional recordsets.

(3) record (Transaction): T and T D, a record in the database.

(4) item set (Itemset): k-itemset (k item set), a set of items that occur at the same time.

(5) support degree (Support): supp (X) = occur (X) / count (D) = P (X).

(6) confidence level (Confidence/Strength): conf (X->Y) = supp (X Y Y) / supp (X)

= P (Y|X). (2) all itemsets (Items): I, the collection of all items.

(7) candidate set (Candidate itemset): C[k], the item set obtained by downwards merging.

(8) frequent itemsets (Frequent itemset): support is greater than or equal to a specific minimum support (Minimum).

The item set of Support/minsup). It is represented as L[k]. Note that the subset of frequent sets must be frequent sets.

(9) lifting ratio (Lift): lift (X > Y) = lift (Y > X) = conf (X >).

Y) /supp (Y) = conf (Y > X) /supp (X) = P (X and Y) / ((()) ()). Note: After the analysis of association rules, the higher the ratio of selling (according to a rule) to blindly selling (generally speaking, the whole data), the better. We call this rule strong rules.

(10) Pruning step: A frequent set is a frequent set only if the subset is a candidate of a frequent set, and the filtering process is a pruning step[3]

3.3.3 The core process of Apriori algorithm based on Apriori algorithm is as follows:

(1) Calculate the support degree of each item set through single scan database D, and get the set of frequent item set.

(2) Connection step: In order to generate, pre-generated, two sets of frequencies with only one item belonging to a different set do a (k-2) JOIN operation.

(3) pruning step: as a result of the superset, some elements may not be frequent. If a subset of a potential K itemset is not a member, the potential frequent itemset cannot be frequently removed from it.

(4) Through a single scan of the database D, the support of each item set is calculated, and the items that do not meet the support are removed to form an iteration loop, repeating steps 2-4 until a certain R value makes it empty, then the algorithm stops.[4]

4. Experiment and result analysis

The algorithm is programmed by arules software package in R language. The core function of the package is Apriori (), and the Apriori algorithm can be implemented. The basic format of the function is Apriori (data, parameter = NULL, apprearance = NULL, control = NULL), in which the parameter of parameter can support and confidence (Confifidence) Dence0, maxlen / minlen of the number of items contained in each itemset, set with important parameters such as output target. The parameter appearance can restrict the prerequisites X and the specific items contained in the association result Y. The Control parameter is used to control the performance of the function, such as setting up the itemsets.(sort=1) or descending order (sort=-1), whether to report the process to the user (VER) Bose=FALSE/TRUE) and so on.[5]

4.1 Data sources

In this paper, 1000 samples of a genetic disease (hereditary disease A) were collected, which included disease information of 1000 samples, coding information of 9445 loci in the samples, and gene information containing these loci. The information for genetic disease A, a column of 0 and 1, contains 500 0,500 1, indicating that we now have a total of 1,000 samples, 500 0 of which are 500 people without disease A and 500 1 of which are 500 people with genetic disease A.

The above 1000 samples contain all the locus information on a chromosome segment. There are 1001 rows and 9445 columns. Specifically, the first line represents the name of the 9445 loci, all beginning with the letter rs; then, there are 1000 lines, each representing a sample of all the loci (9445) on that segment of the chromosome.Each dat file contains the names of several loci representing the locus information contained in the gene, which in fact can be understood as a collection of several loci. Notice that the genotype. dat file contains the coding information for all the loci, so we can get the coding information for each gene.

4.2 Data extraction and processing

In R language, the data is first transformed into an evacuation matrix, and then the evacuation matrix is transformed into a data structure available to Apriori functions. When using association algorithm, we first try to observe the output of Apriori function with the least restriction, and then decide the next step. Here, the minimum support threshold (minsup) is set to 0.5, the minimum confidence threshold (mincon) is set to 0.6, other parameters are not set to default values, and the resulting association rule name is rules0, and the inspect function is used to display the detailed Association rules. In the following results, the LHS column is the left of the association rule. On the side, the RHS column is the right side of the association rule, and support, confidence, lift are support, trust, and promotion, respectively. In the process of parameter co-adjustment, if we pay more attention to the proportion of related items in the whole, we can improve the support degree appropriately. If we pay more attention to the reliability of the rules, we can increase some confidence values. The degree of elevation represents the ratio of the possibility of containing both Y and Y under the condition of containing X to the possibility of containing Y in the itemset without this condition, that is, on the basis of the possibility of Y itself appearing P (Y), the degree of elevation of X to the "mirror rate" P (Y | X) of Y: the index and the degree of confidence are also used to measure the rule's feasibility. Reliability can be regarded as a complementary index of confidence. When lift value is 1, X and Y are independent of each other, X has no effect on the possibility of Y, and the higher the lift value (> 1) indicates that X has a greater effect on Y, that is, the stronger the correlation.[6] Be-

cause of the above theoretical knowledge, we know that lifting degree can be regarded as the most reliable index of association rules, and the conclusions are useful. So when setting parameters, we select the most useful association rules according to lifting only ascending order.

4.3 Pathogenic analysis of single locus, select support = 0.2, confidence = 0.5, in order to minimize missed selection, in descending order according to lift value, take lift > 1.1 all the rules. The results found qualified rules.

The 23 is as shown in Table 2.

genotype	lhs=>rhs	support	confidence	lift
rs2273298	{m=AA}=>{V1=YES}	0.305	0.578747628	1.157495256
rs7543405	{m=AA}=>{V2=YES}	0.278	0.566191446	1.132382892
rs2143810	{m=AA}=>{V3=YES}	0.256	0.560175055	1.120350109
rs7543486	{m=AA}=>{V4=YES}	0.216	0.559585492	1.119170984
rs10779765	{m=AA}=>{V5=YES}	0.217	0.559278351	1.118556701
rs4646092	{m=AA}=>{V6=YES}	0.272	0.557377049	1.114754098
rs1201394	{m=AA}=>{V7=YES}	0.204	0.557377049	1.114754098
rs2244300	{m=AA}=>{V8=YES}	0.221	0.556675063	1.113350126
rs10864304	{m=AA}=>{V9=YES}	0.232	0.556354916	1.112709832
rs4949238	{m=AA}=>{V10=YES}	0.211	0.555263158	1.110526316
rs12128558	{m=AA}=>{V11=YES}	0.203	0.554644809	1.109289617
rs4845881	{m=AA}=>{V12=YES}	0.219	0.55443038	1.108860759
rs9659647	{m=AA}=>{V13=YES}	0.23	0.554216867	1.108433735
rs1257163	{m=AA}=>{V14=YES}	0.207	0.553475936	1.106951872
rs1541318	{m=AA}=>{V15=YES}	0.245	0.553047404	1.106094808
rs873319	{m=AA}=>{V16=YES}	0.201	0.552197802	1.104395604
rs946758	{m=AA}=>{V17=YES}	0.25	0.55187638	1.103752759
rs11121557	{m=AA}=>{V18=YES}	0.267	0.551652893	1.103305785
rs10864301	{m=AA}=>{V19=YES}	0.235	0.551643192	1.103286385
rs6541003	{m=AA}=>{V20=YES}	0.204	0.551351351	1.102702703
rs1033867	{m=AA}=>{V21=YES}	0.207	0.550531915	1.10106383
rs277671	{m=AA}=>{V22=YES}	0.23	0.550239234	1.100478469
rs28603108	{m=AA}=>{V23=YES}	0.231	0.55	1.1

Table 2. From the result analysis, we can see that the sites with strong correlation with disease A are strong to weak, as shown in the above table.

4.4 Single gene (multiple loci) pathogenic analysis experiments

Select support=0.015, confidence=0.95, and the results are shown in Table 3.

able 3 experimental results of single gene (multiple loci)

{1=CC}	=>	{V=YES}	0.026	0.604651163	1.209302326	4
{0=AG,1=CC}	=>	{V=YES}	0.015	0.789473684	1.578947368	4
{0=AA,1=TC}	=>	{V=YES}	0.024	0.615384615	1.230769231	4

Table 3. The results showed that the No. 4 gene was in accordance with the rules, and the four gene was associated with A.

5. Conclusion

Association rule analysis method can effectively discover the complex relationship between disease and gene sites, in order to find the genetic disease-related gene pathogenic sites. In this paper, an Apriori-based genetic disease gene analysis method is proposed and implemented in R language. The experimental results show that the method can effectively detect the genetic pathogenic loci of genetic diseases.

References

1. Ouxin, Shi Lisong, Wang Fan, Wang Qing. Progress and Reflection on Genome-wide Association Analysis [J]. Progress in Physiology, 2010, 02: 87-94.
2. Luo Jinling. Research and Development of Intelligent Instruction System Based on Data Mining Technology [D]. University of Electronic Science and Technology, 2012.
3. jingwhale. http://www.cnblogs.com/jingwhale/p/4618351.html. 2015, 7, 03.
4. learning is persis tent. Http://blog.csdn.net/ learning da/ article/ details/ 51194312. 2016.04.19.
5. Li Yuqi, Wang Xiaoling.Data mining model based on improved Apriori algorithm and genetic algorithm[J].Computer and Digital Engineering, 2007, 03: 16-18+199.
6. Implementation of Liu Jingyi and Zhu Guiling.R Language in Apriori [J]. Scientific Chinese, 2016, 23: 45.
7. Hou Yajun. R language in data mining application [J]. Journal of Jincheng Institute of Technology, 2014,02: 63-65.
8. Data Mining Application of Chen Rongxin.R Software [J]. Journal of Chongqing Business University (Natural Science Edition), 2011, 06: 602-607.
9. Development and Application of Xiao Yingwei, Ge Ming.R Language in Data Preprocessing [J].Journal of Hangzhou University of Electronic Science and Technology, 2012, 06: 165-168.
10. Zou Liling, Zhao Nieqing, Qin Guoyou, Qian Ji, Shao Minhua. Association rules were used to screen SNP loci and their combinations for disease-related diseases [J]. China Health Statistics, 2009, 03: 226-228+233.
11. Sun Hongxia, Du Weinan, Fonford. Statistical analysis methods commonly used in gene mapping of human diseases[J]. Journal of Chinese Academy of Medical Sciences, 2001, 01: 86-88+96.
12. Sun Yulin, Liu Fei, Zhao Xiaohang. Genome-wide Association Analysis of Copy Number Variation[J].Advances in Biochemistry and Biophysics, 2009, 08: 968-977.
13. G.Sangeetha,Sornamaheswari. Density Conscious Subspace Clustering for High Dimensional Data using Genetic Algorithms[J]. International Journal of Computer Applications, 2010, 104.
14. Philip S. Yu. Review - Mining Association Rules between Sets of Items in Large Databases[J]. ACM SIGMOD Digital Review, 1999, 1.